UTILIZACIÓN DE MODELOS DE INTERPOLACIÓN DE MARKOV PARA LA IDENTIFICACIÓN DE SECUENCIAS DE GENES

Marco Gerardo Torres Andrade*

Resumen

Uno de los principales problemas a resolver en la bioinformática es la identificación de genes dentro de un código genético. Una de las alternativas que presenta mayor aplicabilidad en este tipo de problemas es el uso de modelos basados en cadenas de Markov; sin embargo, existe un conjunto de modelos conocidos como modelos de interpolación de Markov (IMM) que presentan un mejor desempeño para este tipo de problemas. Los IMM son utilizados por GLIMMER en la identificación de genes con bastante éxito.

Palabras clave: Bioinformática, Secuenciamiento, Identificación de genes

Abstract

One of most important problems in bioinformatics consists on the identification of genes in a genome. An alternative used for obtaining this objective consists of using applications based on Markov Chains. However, there are a set of models known as Interpolated Markov Models (IMM) that present best results. This kind of models are used by GLIMMER with excelent results.

Key Words: Bioinformática, Secuenciamiento, Identification of genes

* Ingeniero Electricista de la Universidad Nacional de Colombia. Actualmente adelanta sus estudios de magíster en Ingeniería de Sistemas en la Universidad Nacional de Colombia. Docente de tiempo completo de la Universidad de Cundinamarca, sede Fusagasuga. **marco_gerardo@hotmail.com**

1. Introducción

El problema de realizar una identificación precisa de los genes en genomas microbianos ha ido creciendo en importancia en la medida en que han incrementado la cantidad de proyectos relacionados con el secuenciamiento de genomas.

Existe una gran cantidad de datos de genomas microbianos dispuestos para ser analizados; por ello se ha ido requiriendo de sistemas computacionales más precisos que permitan explorar los genomas y aportar nuevos conocimientos sobre ellos.

Uno de los primeros pasos en el análisis del genoma microbiano consiste en identificar todos sus genes. Este tipo de genomas tienden a ser muy ricos en contenido genético; poseen del orden del 90% de la secuencia codificada; el trabajo es más complicado en células eucarióticas, pues muchas poseen hasta menos del 10% de la secuencia codificada [1].

La razón es la siguiente: Los ADN bacterianos no tienen intrones, razón por la cuál la proteína es colineal con el ADN. En organismos superiores, las proteínas no son colineales con el ADN debido al procesamiento recibido por el hnARN, en el cuál son cortados los intrones, para producir el mRNA que va a ser utilizado para la traducción. De ahí radica el problema de los genes a partir de la anotación del genoma. Una de las anotaciones más importantes consiste en determinar cuáles de las secuencias del genoma son genes. Debido a la no colinealidad producida en las células eucarióticas, de la cuál se hizo mención anteriormente, es necesario para la detección de genes definir algo que se conoce como marcos abiertos de lectura (ORF), Los cuales son segmentos de DNA que comienzan en una secuencia de iniciación ATG (metionina) y en marco encuentra una secuencia equivalente a un codón de terminación [2].

El problema más difícil es determinar cuáles de dos o más marcos abiertos de lectura representan genes verdaderos. Otra dificultad que se presenta consiste en identificar el inicio de la traducción y encontrar señales reguladoras tales como los promotores y las secuencias terminadoras.

La forma más confiable de identificar un gene en un nuevo genoma es encontrar un homólogo cerrado de otro organismo. Ello puede hacerse de forma efectiva mediante el uso de herramientas como el BLAST y el FASTA [1]. Sin embargo, muchos de los genes en genomas nuevos no encuentran un homólogo significativo en genes conocidos. Para este tipo de genes se hace necesaria la utilización de algún método computacional que realice la valoración de la región codificada para evaluar los genes. El programa más conocido para tal fin es el GeneMark [1], el cuál usa cadenas de Markov para valorar regiones codificadas.

2. Cadenas de Markov

Una cadena de Markov es una secuencia de variables aleatorias X(p) donde la distribución de probabilidad de cada X(i) depende de una cantidad de k valores precedentes X(i-1), X(i-2),..,X(i-k) [3].

Cada uno de los nucleótidos de una secuencia puede diferenciarse por su base correspondiente; por tal motivo es usual hablar indistintamente de nucleótido o de base. Los modelos de cadenas de Markov aplicados a secuencias de DNA permiten calcular la probabilidad de ocurrencia de un nucleótido (es decir, de la base correspondiente a un nucleótido dado) b en la secuencia dependiendo de los k nucleótidos inmediatamente anteriores a b en la secuencia.

Los modelos de Markov se utilizan bastante en el análisis de secuencias de datos biológicos. Los más utilizados son las cadenas de Markov de orden fijo, en las cuales el contexto (es decir, la secuencia de k

nucleótidos precedentes) se utiliza en cada posición donde se desea hallar la probabilidad de ocurrencia de un nucleótido dado.

Cualquier modelo de cadenas de Markov de orden fijo predice un nucleótido de la secuencia de DNA usando los nucleótidos anteriores de la secuencia; el modelo de mayor uso es el de cadenas de Markov de quinto orden.

3. Modelos de interpolación de Markov (IMM)

Además existe una herramienta denominada GLIMMER, la cuál usa una técnica denominada Modelos Interpolados de Markov (IMM) para encontrar regiones codificadas en secuencias microbianas [4].

Los IMM se pueden considerar en principio más fuertes que las cadenas de Markov. Varios experimentos realizados con esta herramienta han demostrado que produce resultados más precisos al ser usado para encontrar genes en DNA bacterianos [4].

4. Ventaja que presentan los IMM sobre las cadenas de Markov

El inconveniente que presentan los modelos de cadenas de Markov es que el aprendizaje de algunos modelos puede producir dificultad cuando no se tienen suficientes datos de entrenamiento para que el sistema pueda hacer una estimación precisa de la existencia de una base determinada con cada combinación de bases precedentes. Para poder realizar, en estos casos, una estimación adecuada, deben presentarse una cantidad muy grande de ocurrencias de todas las posibles combinaciones.

Los sistemas basados en IMM funcionan de forma más adecuada para este tipo de problemas, pues utiliza la combinación de probabilidades para contextos de diferentes longitudes de nucleótidos

precedentes, y solo mediante el uso de estos contextos (oligómeros) se puede suministrar una cantidad de datos suficiente para el entrenamiento del modelo [4].

5. Forma en que se usan los IMM

Un IMM utiliza una combinación lineal de probabilidades obtenidas para diferentes longitudes de oligómeros que permiten realizar mejores predicciones, ya que el sistema le otorga una alta valuación a los oligómeros que presentan una alta ocurrencia, mientras que a aquellos que son poco comunes reciben un valor bajo.

Así es como los IMM pueden utilizar un contexto largo y realizar a partir de él una predicción lo más precisa posible, haciendo uso de las grandes ventajas en cuanto a la precisión que poseen los modelos de Markov de orden superior.

Conclusiones

1. El interés creciente que ha producido el secuenciamiento de genomas ha hecho que la informática aporte elementos cada vez más valiosos para la detección de genes.

2. En genomas bacterianos existe gran parte de su código secuenciado, ya que sus proteínas son colineales con el ADN; en cambio, en células eucarióticas, el secuenciamiento se ha constituido en una tarea más difícil debido a que, por el corte de los intrones en el procesamiento que recibe el hnARN para conformar el mARN que es traducido posteriormente, la secuencia se halla fragmentada en el genoma. De ahí que una de las anotaciones del genoma de interés para la ciencia ha consistido en detectar cuáles secuencias son genes.

3. Para ello es necesario definir marcos abiertos de lectura (ORF); estos son segmentos de DNA que comienzan en una secuencia de iniciación y en marco terminan en un codón de terminación.

4. La dificultad consiste en determinar cuales ORF corresponden a genes verdaderos, así como también la determinación de promotores, secuencias de terminación y otro tipo de señales de interés.

5. Una solución confiable consiste en encontrar homólogos cerrados del organismo que se esté estudiando; sin embargo, para algunos de ellos no es fácil obtener homólogos cerrados apropiados.

6. Es por este motivo que el problema de detección de genes ha entrado a ser estudiado por informática, proponiendo modelos estocásticos que permiten determinar la probabilidad de ocurrencia de un nucleótido dado a partir de su contexto, es decir, la secuencia de bases precedentes. Un modelo bastante usado es el de las cadenas de Markov. Una cadena de Markov es una secuencia de variables aleatorias X(p) donde la distribución de probabilidad de cada X(i) depende de una cantidad de k valores precedentes X(i-1), X(i-2),..,X(i-k). Un modelo de cadena de Markov ampliamente usado para detectar genes es el Modelo de Markov de quinto orden, el cuál utiliza un contexto de 5 bases para hacer la predicción.

7. El inconveniente que presentan los modelos de cadenas de Markov es que el aprendizaje de algunos modelos puede producir dificultad cuando no se tienen suficientes datos de un contexto determinado. De ahí que haya surgido un modelo que presenta mayores ventajas, llamado modelo de interpolación de Markov (IMM), el cuál utiliza una combinación lineal de probabilidades obtenidas para diferentes longitudes de oligómeros que permiten realizar mejores predicciones, valorando con una alta calificación a aquellos que presentan una alta ocurrencia y con una puntuación baja a los de ocurrencia

poco frecuente. Esta fortaleza la combina con las ventajas en precisión que tienen los modelos de cadenas de Markov, lo cuál hace de esta técnica bastante eficiente para la identificación de genes.

Referencias Bibliográficas

[1] DELCHER, Arthur L et.al. (1998). Microbial gene identification using interpolated Markov models. Oxford University press, Nucleic Acids Research, vol 26, No 2.

[2] DELCHER, Arthur L et.al. (1999). Improved microbial gene identification with GLIMMER. Oxford University press, Nucleic Acids Research, vol 27, No 23

[3] MOJICA, Tobías et.al. (2001). El lenguaje genético. Capítulo 8: Tecnologías moleculares para analizar los genes. Editorial Celsus.

[4] Ross, Sheldon. (1996). *Stochastic processes.* John Willey and Sons.